LAMP-TR-148 February 2008
HCIL-2008-07

How Do Users Find Things with PubMed? Towards Automatic Utility Evaluation with User Simulations

Jimmy Lin[1] and Mark D. Smucker[2]

[1]The iSchool
University of Maryland
jimmylin@umd.edu

[2]Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts Amherst
smucker@cs.umass.edu

Abstract

In the context of document retrieval in the biomedical domain, this paper explores the complex relationship between the quality of initial query results and the overall utility of an interactive system. We demonstrate that a content-similarity browsing tool can compensate for poor retrieval results, and that the relationship between retrieval performance and overall utility is non-linear. Arguments are advanced with user simulations, which characterize the relevance of documents that a user might encounter with different browsing strategies. With broader implications to IR, this work provides a case study of how user simulations can be exploited as a formative tool for automatic utility evaluation. Simulation-based studies provide researchers with an additional evaluation tool to complement interactive and Cranfield-style experiments.

Publication Date: February 5, 2008
Keywords: Related article search, find-similar

Please cite as: Jimmy Lin and Mark D. Smucker. How Do Users Find Things with PubMed? Towards Automatic Utility Evaluation with User Simulations. Technical Report LAMP-TR-148/HCIL-2008-07, University of Maryland, College Park, February 2008.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
FEB 2008		**00-00-2008 to 00-00-2008**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
How Do Users Find Things with PubMed? Towards Automatic Utility Evaluation with User Simulations	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Maryland,Human Computer Interaction Lab,College Park,MD,20742**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

In the context of document retrieval in the biomedical domain, this paper explores the complex relationship between the quality of initial query results and the overall utility of an interactive system. We demonstrate that a content-similarity browsing tool can compensate for poor retrieval results, and that the relationship between retrieval performance and overall utility is non-linear. Arguments are advanced with user simulations, which characterize the relevance of documents that a user might encounter with different browsing strategies. With broader implications to IR, this work provides a case study of how user simulations can be exploited as a formative tool for automatic utility evaluation. Simulation-based studies provide researchers with an additional evaluation tool to complement interactive and Cranfield-style experiments.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**17**	

1 Introduction

This work was motivated by a simple question: "How do users find things with PubMed?" PubMed is a large, publicly-accessible Web-based search engine that provides access to MEDLINE, the authoritative repository of abstracts from the medical and biomedical primary literature. Both are maintained by the U.S. National Library of Medicine (NLM). MEDLINE currently contains over 17 million abstracts, covering a wide range of disciplines within the health sciences (broadly interpreted), from biochemistry to public health. As the primary access point to MEDLINE, PubMed is an indispensable tool for clinicians and scientists.

There is substantial evidence to suggest that PubMed is difficult to use. At the core, it is a pure Boolean retrieval engine that returns results sorted in reverse chronological order. A number of studies have demonstrated the superiority of ranked retrieval over comparable Boolean techniques (e.g., [25]). Since almost all commercial Web search engines implement some sort of best-match algorithm, users have grown accustomed to using ranked retrieval systems. In contrast, the query formulation process in PubMed feels quite foreign. Empirical support comes from PubMed transaction logs, where long chains of repeated query formulations are frequently found—they suggest that users often struggle coming up with the right query terms.[1] In fact, approximately a fifth of all PubMed queries return zero results. Related to these challenges is the difficulty associated with controlling the result set size, which is another characteristic typical of Boolean retrieval. For example, adding an additional term to a query that retrieves 1000 hits might yield zero hits.

The importance of access to the primary literature for clinicians and scientists lends merit to our motivating question. Abstracting from this specific instance into a more general problem, we explore the complex relationship between the quality of initial query results and the overall utility of an interactive system. In particular, we examine the contributions of a browsing tool based on content similarity: we hypothesize that such a feature can compensate for poor retrieval results. Through strategy simulations, we uncover a non-linear relationship between utility and quality of the initial results. These findings highlight the need to evaluate an interactive IR application as a whole, and not simply component-wise.

More generally, this work represents a case study in the application of user simulations to automatically measure utility. We argue that simulation-based evaluations provide researchers with an alternative to existing methodologies, as well as a powerful formative tool that combines advantages of both interactive and Cranfield-style evaluations.

2 Background

Given that query formulation in PubMed is a major issue, how do users go about finding relevant documents? We believe that related article suggestions contribute significantly to a user's overall search experience. This section overviews this feature and discusses our hypothesis.

Whenever the user examines an abstract in PubMed, the right panel of the browser is automatically populated with titles of articles that may also be of interest, as determined by a probabilistic content-similarity algorithm [19] (see Figure 1). That is, each abstract view triggers a related article search: the top five results are integrated into a "Related Links" panel in the display.[2] This feature is similar to what Smucker and Allan [22] call *find-similar*; cf. [10, 28]. Related article suggestions provide an effective browsing tool for PubMed users, allowing them to navigate the document collection without explicitly issuing queries.

[1] Jimmy Lin and W. John Wilbur, paper under review.

[2] Although MEDLINE records contain only abstract text, it is not inaccurate to speak of searching for articles since PubMed provides access to the full text when available; we use "document" and "article" interchangeably in this paper.

Figure 1: Typical PubMed screenshot showing a MEDLINE abstract and "Related Links".

We hypothesize that related article suggestions compensate for cases where PubMed results are poor (whether due to difficulties in query formulation or lack of relevance ranking). As long as the initial results contain one relevant document, related article suggestions can help the user find more relevant documents by browsing around. Thus, the quality of the initial results is only one factor affecting the overall utility of the system.

Prima facie support for this argument comes from the cluster hypothesis [26]. Since relevant documents tend to be clustered together (i.e., similar in content), a browsing tool based on content similarity should be effective in helping users gather relevant documents. This bulk of this paper focuses on simulations of search strategies, which appear to support this claim. However, we also relate experimental findings to a recent analysis of transaction logs from PubMed, lending further credibility to our conclusions (see Section 7).

3 User Simulations

IR evaluations generally fall into one of two categories: batch-style, system-centered evaluations in the Cranfield tradition [6] (as exemplified by *ad hoc* evaluations in TREC), and interactive, user-centered evaluations of searcher performance ([15] provides a classic example). Researchers have long acknowledged that Cranfield-style evaluations are limited in examining only one aspect of information seeking—the quality of a ranked list generated by a one-shot query. To measure the utility of interactive retrieval systems, one must typically turn to carefully-orchestrated user studies that examine human search behavior. Many studies over the years have confirmed that the first is not a substitute for the second [3, 9, 23, 24], since a significantly "better" retrieval algorithm (as measured by batch evaluations) might not lead to significantly better utility. Given these facts, why aren't there more interactive evaluations that focus on utility?

The short answer is that interactive evaluations are difficult to conduct, not for lack of trying [8]. Compared to batch evaluations, the high-cost and time-consuming nature of interactive evaluations limit the speed at which hypotheses can be explored and the statistical significance of results. Due to the conflation of numerous user, task, and contextual factors that are unavoidable, even with the

adoption of best practices in study design, results are difficult to compare and often do not support convincing generalizations.

Simulation-based evaluations have recently emerged as a promising methodology for bridging interactive and batch evaluations [17, 20, 21, 27]; see also similar work in the HCI community [5, 11]. Although details vary, they are all based on a common idea: instead of evaluating real users, simulate what they might do. In other words, the evaluation centers on examining the behavior of an idealized user who adopts a particular (known) strategy for interacting with retrieval systems.

User simulations can be viewed as a compromise between interactive and batch evaluations. They preserve the advantages of batch evaluations in that experiments remain easily and rapidly repeatable, while at the same time they begin to characterize the utility of an interactive system, potentially modeling user, task, and context factors. Naturally, the validity of simulation results is contingent on the realism of the simulations—but it is possible to have meaningful debates over the realism of different user models, informed by results of user studies [10], eye-tracking experiments [13], log analysis [12], etc.

In fact, the Cranfield methodology can be viewed as a primitive user simulation: it models a user who types in a query and then examines the results sequentially. Many of the criticisms leveled against it speak to the poor assumptions it makes about users: one shot retrieval goes against what we know about the iterative nature of information-seeking behavior; the assumption of binary, independent relevance judgments is an over-simplification of the complex nature of relevance; etc. With user simulations, we can begin to address each of these deficiencies in a principled fashion.

One potential concern with simulation-based evaluations is comparability of results. Characterizations of utility depend both on the system and the "simulation module", so the latter must be distributed in the same way that topic/qrels are widely available today. We envision the evolution of standard test suites, of which Cranfield test collections represent one specific type. Naturally, the community as a whole would need to converge on what these standard test suites might contain. We are hopeful that such a consensus is possible—in the same way that mean average precision and other evaluation methods became standard practice after much debate in the early days of IR.

Of course, the emergence of user simulations as an evaluation methodology does not obviate the need for interactive evaluations—there can ultimately be no replacement for users when the goal is to develop systems that are useful for human beings. We advocate simulations as a formative tool, replacing user studies in situations where they are simply too slow or cumbersome (e.g., for rapid prototyping). With the distribution of standard test suites, simulation-based evaluations should be no more difficult to conduct than current Cranfield-style experiments, and hence they represent a superior alternative.[3] In our view, traditional user studies will most likely remain the most effective tool for summative evaluations.

Finally, user simulations might be used prescriptively as well as descriptively. That is, results of user simulations could be used as a basis for educating users on effective search strategies. This is not an unrealistic scenario in the context PubMed: due to the nature of its users and their work, PubMed searchers are often willing to learn effective search techniques and advanced features.[4]

4 Experimental Setup

We began by abstracting pertinent elements of the problem into a more controlled experiment, preserving the overall goal of the study: to characterize the impact of a content-similarity browsing tool on utility.

[3]More accurately, Cranfield-style experiments would be subsumed as one test suite in a simulation-based evaluation model.

[4]Empirical evidence for this claim is demonstrated by the numerous tutorials and mini-courses offered on PubMed, as any casual Web search will reveal.

Information describing standard [methods or protocols] for doing some sort of experiment or procedure. *methods or protocols:* purification of rat IgM
Information describing the role(s) of a [gene] involved in a [disease]. *gene:* PRNP *disease:* Mad Cow Disease
Information describing the role of a [gene] in a specific [biological process]. *gene:* casein kinase II *biological process:* ribosome assembly
Information describing interactions between two or more [genes] in the [function of an organ] or in a [disease]. *genes:* Ret and GDNF *function of an organ:* kidney development
Information describing one or more [mutations] of a given [gene] and its [biological impact or role]. *gene with mutation:* hypocretin receptor 2 *biological impact:* narcolepsy

Table 1: The five templates used in the TREC 2005 genomics track (with sample instantiations).

The general setup was as follows: starting from an initial ranked list in response to an information need, we simulated user behavior under different browsing strategies. Each simulation is characterized by a sequence of documents, which represents the order in which the user would examine documents in the collection. Since both the input and output of the simulation are ordered lists of documents, we can assess and compare their quality using standard ranked retrieval metrics—this is similar to the strategy used by Aalbersberg [1] for evaluating relevance feedback. In what follows, we describe the test collection, the initial results, the simulation procedure, and metrics used to capture utility.

4.1 Test Collection

Evaluations were conducted with data from the TREC 2005 genomics track [7], which employed a ten-year subset of MEDLINE (1994–2003). The collection contains approximately 4.6 million records, or approximately a third of the entire database at the time it was collected in 2004 (commonly known as the MEDLINE04 collection).

One salient feature of this TREC evaluation was its use of generic topic templates (GTTs), which consist of semantic types, such as genes and diseases, embedded in prototypical information needs, as determined from interviews with biologists and other researchers. In total, five templates were developed, each with ten fully-instantiated topics; examples are shown in Table 1.

For each topic, relevance judgments were provided by an undergraduate student and Ph.D. researcher in biology. No relevant documents were found for one topic, which was discarded from our experiments.

4.2 Initial Results

As input to the user simulations, we wished to consider initial results that range widely in terms of quality. Since our hypothesis concerns situations where browsing compensates for poor results, we especially needed samples of such. Note that we are careful not to equate poor results with a poor

retrieval algorithm, since query formulation may play an important role (as in the case of PubMed). In addition, variations in topic difficulty, as well as variations in performance exhibited by even the best retrieval algorithms, contribute to poor query results as well.

Although Turpin and Scholer [24] present a technique for synthetically generating ranked lists that attain a specific mean average precision, we rejected their method since it does not yield results that correspond to any real system. Instead, we used as input all 62 runs submitted to the TREC 2005 genomics track (58 of which contributed to the pool). This gave us an accurate sampling of the types of results generated by modern retrieval engines. For the submitted runs, MAP ranged from 0.302 to 0.054 (mean of 0.197); P10 ranged from 0.474 to 0.176 (mean of 0.358).

4.3 Simulation Procedure

We examined two different content-similarity algorithms and two different browsing strategies, yielding a two-by-two matrix experiment. Much of our procedure is similar to that used by Smucker and Allan [22], to which we refer the reader. Here, we provide only an overview.

One experimental variable was the algorithm for suggesting related articles. We considered two:

- Using language modeling retrieval, implemented with Lemur [16], treating title and abstract as the "query". Settings used: the Krovetz stemmer, Dirichlet prior smoothing with $\mu = 1500$.
- Using the content-similarity algorithm in PubMed, accessed through the PubMed eUtils API.[5]

Quite explicitly, our goal was not to compare Lemur with PubMed, but rather to examine the effects of different content-similarity algorithms, given that the two have different theoretical foundations.

In terms of browsing behaviors, we examined the two proposed by Smucker and Allan [22]: the *greedy* pattern represents an abstraction of depth-first behavior in examining a ranked list and the *breadth-like* pattern represents an abstraction of breadth-first search behavior. Both were adapted from the findings of an eye-tracking study conducted by Klöckner et al. [14], and are consistent with the results of Aula et al. [4]; cf. [13].

Under the greedy strategy, the user starts with the initial ranked list and examines documents in rank order. Whenever a relevant document is encountered, the user applies content-similarity search and pulls up its list of related documents (in the current PubMed interface, this is equivalent to clicking on the "See all Related Articles" link). The user ceases to examine documents in a list after examining 2 contiguous non-relevant documents. After stopping, the user hits the "back button" and returns to the previous list and continues examining documents in that list (unless the user is already examining the initial results, in which case the user simply continues down the list).

Under the breadth-like strategy, the user also examines documents in rank order. Unlike the greedy pattern, the breadth-like browser only begins to examine related article suggestions when the ranked list's quality becomes too poor. As the user examines relevant documents, documents are placed in a first-in first-out queue local to the current list. When the precision at N, where N is the rank of the current document, drops below 0.5 or when 2 contiguous non-relevant documents have been encountered, the user applies content-similarity search to the first relevant document in the queue. When the user returns to the current list, the user applies content-similarity search to the next document in the queue until the queue is empty. The browser never applies content-similarity search on a relevant document more than once. The breadth-like strategy models a user who delays exploration until the current list seems to have gone "cold." The user stops examining a ranked list in the same manner and with the same criteria as the greedy browser, i.e., hitting the "back button" after encountering 2 contiguous non-relevant documents.

[5] http://www.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html

In both simulations, the user never examines a document more than once, even though it is likely that a hit appears in several results (given the tendency for documents to cluster together). We see no good reason why a user would want to re-examine documents previously encountered. In a Web interface, visited links are typically marked (e.g., change in link color), providing a prominent visual cue to help users remember where they have already been.

4.4 Measuring Utility

Since ranked lists serve as inputs to our simulations and their outputs consist of ordered document sequences, it is appropriate to evaluate both using standard ranked retrieval metrics. We argue that measurements made on the simulation outputs quantify the overall utility of the system, since they capture the quality of documents that a user might encounter with the system. Therefore, we are not merely measuring one-shot retrieval effectiveness, but rather the usefulness of the system given a particular usage scenario.

As with most Web search engines, early precision is very important, as studies have shown that users focus primarily on top results [2, 13]. Thus, we settled on P20 as one metric—the cutoff is meaningful since PubMed presents 20 results per page. The downside of precision at fixed cutoff, of course, is its inability to capture recall-oriented performance. On the other hand, we feel that MAP at 1000, the most common ranked retrieval metric, does not accurately characterize utility in our case since typical users are unlikely to examine that many hits. MAP at lower cutoffs is also problematic, since there may be more relevant documents than the size of results list—in which case, a MAP of 1.0 would be impossible to achieve. This causes problems since unequal score ranges make meaningful cross-topic comparisons more difficult. Ultimately, we settled on interpolated precision at recall 0.5, which we feel captures a reasonable tradeoff between precision and recall.

5 Results

For each of the conditions in our two-by-two matrix experiment, we ran a total of 3038 separate trials (62 runs $\times$ 49 topics). In the graphs in Figure 2, each trial is plotted as a point in a scatter graph: the x coordinate represents interpolated precision at recall 0.5 (IPR50) for the initial ranked list (baseline), while the y coordinate represents IPR50 for the simulation (utility). The top graph shows the greedy browsing strategy and the bottom graph the breadth-like browsing strategy, both with Lemur. Plots for the PubMed content-similarity algorithm look similar and are not included for space considerations.

For both plots, points above the line $y = x$ represent instances where the content-similarity browsing tool would help users gather more relevant documents than could be obtained with the ranked list alone. For both strategies, we see that substantial improvements are possible at low baseline performance levels—this appears to confirm our hypothesis that a content-similarity browsing tool can compensate for poor query results. As a note, points on the left edge of the plots ($x = 0$) are simply artifacts of the interpolation, since our simulations cannot possible improve initial results containing zero relevant documents.

It is interesting to note that with good initial results, browsing related article suggestions can actually be detrimental, particularly with the greedy browsing strategy, which exploits content-similarity search at the earliest possible moment without regard to the quality of the current results list. Note that in the breadth-like browsing strategy, the user will not even consider related article suggestions if the initial results are good enough (precision above 0.5, no two contiguous non-relevant documents). Thus we see the linear relationship between baseline and utility at high IPR50 values. In general, we are not particularly concerned with the cases of decreased performance because real users generally exhibit some type of lightweight lookahead behavior, which is not captured in our simulations. Eye-tracking studies have shown that before clicking on a link, users often "look ahead" a couple of hits to see if

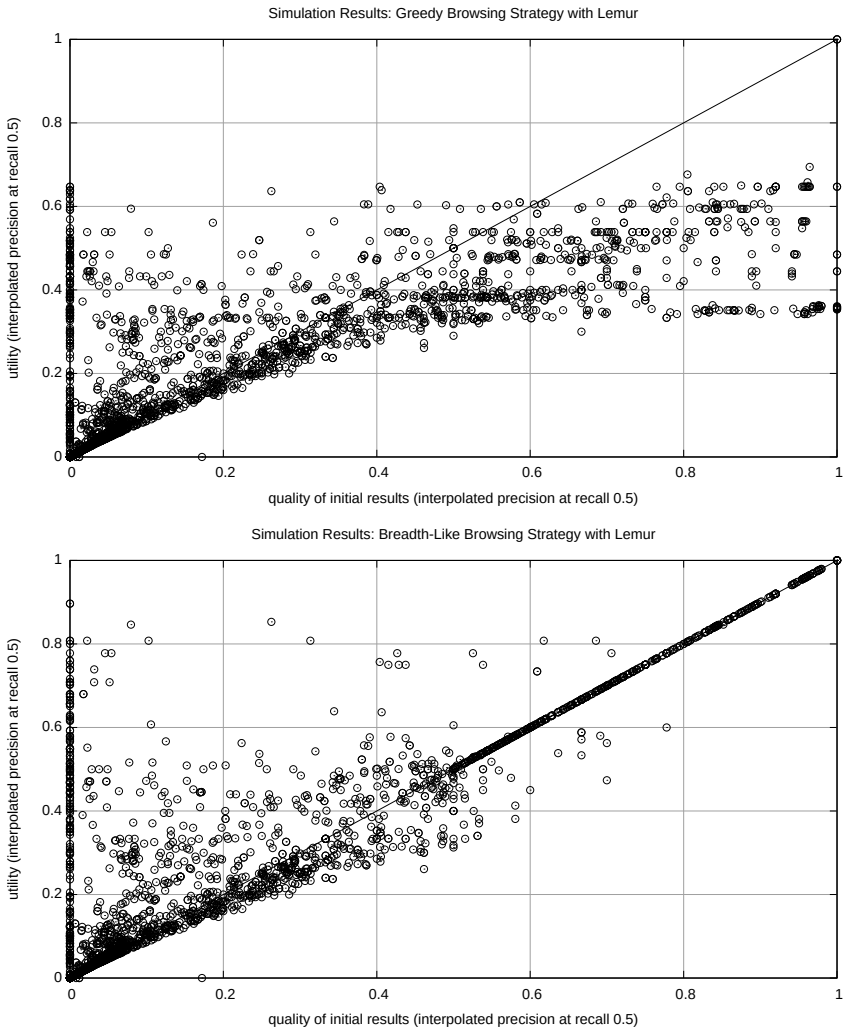

Figure 2: Scatter plots relating quality of initial results to utility (interpolated precision at recall 0.5).

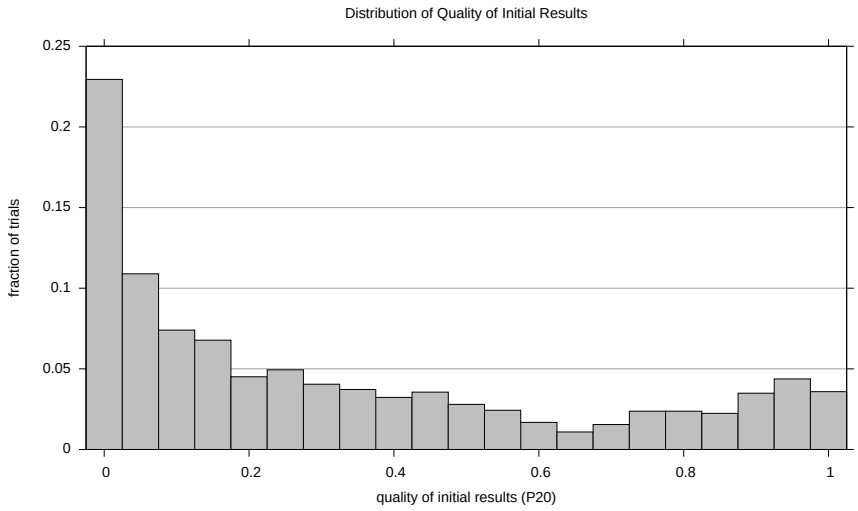

Figure 3: Distribution of P20 scores for the initial ranked lists across 3038 trials (62 runs × 49 topics).

they might be relevant [13]. We believe that this type of behavior would suppress browsing in cases where the initial results were very good. Furthermore, we find that initial results are more often poor than they are good (more on this below).

Since there are only 21 possible P20 values, the metric supports aggregation in a straightforward manner and requires different types of graphs for reporting experimental results. In Figure 3, we show the distribution of P20 scores for all initial results (3038 trials). There appears to be a bimodal distribution, corresponding to "easy" topics (i.e., high P20 scores) and "hard" topics (i.e., low P20 score). For 65% of the trials, the P20 score of the initial results is 0.35 or lower. This underscores the importance of improving bad results and further lessens the impact of cases where the browsing tool might be detrimental (see above).

We see from Figure 3 that in 23% of all cases, no relevant documents were found in the top twenty results, which indicates that on the whole the topics were quite challenging for modern retrieval systems. Note that the content-similarity browsing tool cannot improve on cases where no relevant documents are found in the initial results, since the underlying premise of the tool's effectiveness is that relevant documents cluster together (in our simulations, users only apply content-similarity search to relevant documents).

In Figure 4, we compute the mean P20 utility at each baseline P20 value, showing the results as line graphs: the top figure shows simulations with Lemur as the content-similarity algorithm, while the bottom graph shows simulations with the PubMed algorithm. Both graphs focus on the 65% of cases where P20 is 0.35 or lower.

These results support our hypothesis that a browsing tool based on content similarity can compensate for poor query results. As an example, starting from a ranked list with 15% precision, a user stands to gain about eight absolute percentage points on average from browsing related article suggestions generated by Lemur. Although obviously exaggerated, this translates into a relative improvement of approximately fifty percent! We see similarly large improvements with the PubMed content-similarity algorithm. The fact that consistent gains are achieved, independent of the browsing strategy and content-similarity algorithm suggests that our results cannot merely be artifacts of experimental design.

The downside of the graphs in Figure 3 is that they hide per-topic variations in simulated utility scores. A more detailed analysis is shown in Figure 5, where we graph the fraction of trials at each baseline P20 level that resulted in increases or decreases in simulated P20 utility. Lighter bars represent

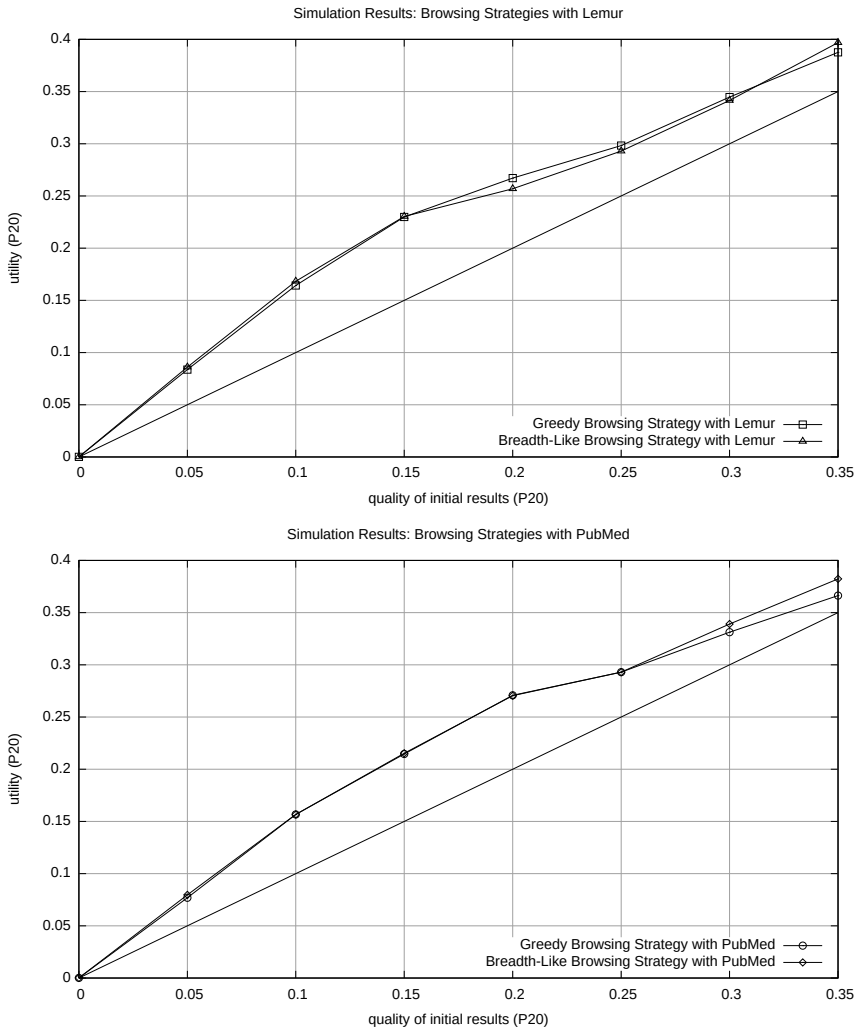

Figure 4: Line graphs relating quality of initial results to utility (P20).

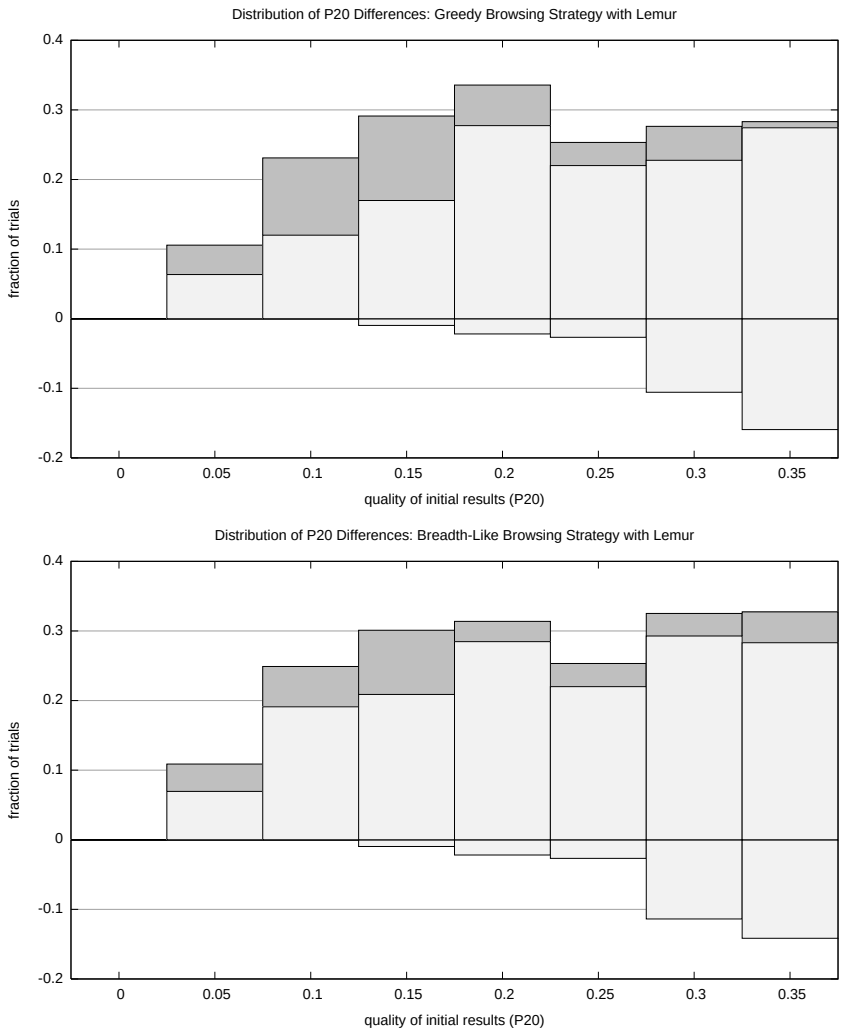

Figure 5: Fraction of trials that lead to increases or decreases in precision at each baseline P20 value. Lighter bars represent $\pm[0.10, 0.30]$, darker bars represent $\pm \geq 0.35$

absolute increases or decreases of between 0.10 and 0.30 (inclusive); darker bars represent absolute increases or decreases of 0.35 or more (although no such decreases were observed). To reduce noise we ignored cases where P20 changed only slightly (± 0.05). The top graph shows results for the greedy strategy with Lemur and the bottom graph shows results for the breadth-like strategy with Lemur (results with PubMed appear similar). As an example of how to understand these graphs: for the breadth-like strategy with Lemur (right graph), of all trials whose initial result set contained 2 relevant documents (P20 of 0.1), about a quarter of the time the simulated user encountered at least two additional relevant documents (P20 gain of at least 0.1), and nearly six percent of the time the user encountered at least seven addition relevant documents (P20 gain of at least 0.35). At that particular baseline P20 level, in no cases did content-similarity browsing hurt.

We see from these results that at baseline P20 scores below 0.25, browsing related article suggestions rarely hurts (less than five percent of the time). These bar graphs further underscore our finding that content-similarity browsing is especially helpful when initial results are poor.

Another significant finding from these experiments is the non-linear relationship between initial result quality and utility. We can divide the utility curves in Figure 4 into two general regions:

- In one region, there is an "amplification effect" with better initial results. That is, utility gains from marginal improvements in initial result quality are larger than expected. This occurs roughly in the P20 intervals [0.05, 0.15] (for Lemur) and [0.05, 0.20] (for PubMed).
- In another region, there is a "diminishing returns effect" with better query results. That is, utility gains from marginal improvements in query result quality are smaller than expected. For example, consider the PubMed results (Figure 4, right): baseline P20 of 0.2 yields 0.27 utility, while baseline P20 of 0.25 yields only 0.29 utility. This occurs roughly in the P20 intervals [0.15, 0.25] (for Lemur) and [0.20, 0.25] (for PubMed).

Similar non-linear relationships between result quality and utility have previously been noted [3, 24]. These observations hold important implications for work on retrieval algorithms or any attempt to improve retrieval performance: the marginal gain in utility depends not only on the marginal improvement in quality of retrieved results, but also on the absolute quality itself. At very low precision levels, the marginal gain in utility is quite high. For example, compare a system that returns one relevant document to a system that returns none; to the extent that the cluster hypothesis holds, a user could browse the document collection to find more relevant documents starting with one relevant document. At other precision levels, the marginal gain in utility could be quite low. For example, more relevant documents returned in the retrieved results would have been found by browsing anyway, thereby lessening the impact of improvements in one-shot retrieval effectiveness. These observations suggest that emphasis in research should be placed on improving poor results, and that averaging per-topic scores hides important variations in performance.

6 Advantages of Simulations

To recap the results thus far: we hypothesized that in an interactive retrieval application, a content-similarity browsing tool can compensate for poor quality results returned by the system. User simulations appear to support this claim, regardless of the actual content-similarity algorithm or specifics of the browsing strategy. In addition to this finding, we also intend for this work to serve as a case study for simulation-based evaluations. We discuss a number of insights that would not have been possible with either a user study or a Cranfield-style experiment:

User studies at scales comparable to our simulations are not practical—our two-by-two matrix design contains 3038 trials for each condition. Needless to say, a large number of trials makes trends much

more apparent and facilitates quantitative comparisons. Furthermore, the simulations have taught us valuable lessons about the experimental design space: for example, that the specific content-similarity algorithm is not as important as one might think. Without this knowledge, it would be natural for a researcher to explore different algorithms in a user study, thereby making less effective use of subjects' time. This represents a great example of leveraging simulations in a formative capacity, to inform the design of summative evaluations.

The simulation module aside (which can be developed into a reusable evaluation component), our evaluations are no more difficult to conduct than traditional Cranfield-style experiments (runs can be rapidly conducted and repeated as often as necessary). Yet, the simulations reveal insights that are not possible in the standard Cranfield methodology. Simulations begin to characterize utility—not only a system's one-shot retrieval effectiveness, but also ways in which a user would search the system. In this specific study, we uncovered a complex relationship between initial retrieval quality and utility, which can be helpful in guiding the development of future retrieval systems.

7 PubMed Results

Let us now return to our original motivating question: "How do users find things with PubMed?" Instead of directly addressing the question, we abstracted a more general problem and explored the effectiveness of a content-similarity browsing tool. In this section, we discuss how experimental findings directly relate to PubMed.

We are able to demonstrate that browsing related article suggestions is an effective information-seeking strategy. This, however, says nothing about whether users actually take advantage of the PubMed feature. Fortunately, independent evidence provides a more complete characterization of user behavior. A recent analysis of PubMed query logs indicates that searchers click on suggested article titles with significant frequency [18]. Data gathered during a one week period in June 2007 indicate that approximately 5% of page views in non-trivial user sessions (discarding, for example, sessions that consist of one page view) are generated from users clicking on related article links. Approximately one fifth of all non-trivial user sessions involve at least one click on a related article link. Furthermore, there is evidence of sustained browsing using this feature: the most frequent action following a click on a related article is another click on a related article (about 40% of the time). Thus, browsing related articles appears to be an integral part of PubMed searchers' activities.

As a final experiment, we ran our strategy simulations on actual PubMed results. One of the co-authors manually formulated PubMed queries interactively for each of the topics in the TREC 2005 genomics track test collection (49 topics). Using the topic templates as a starting point, the "pearl growing" strategy was adopted: the aim was to find a relevant document, and then use it as a basis for learning more about the topic (for query expansion, refinement, etc.). Although the co-author was not a domain expert, the relevance judgments made the task manageable. Approximately 5–10 minutes was spent on each topic: queries were reformulated until at least one relevant document was found in the top 1000 results, or until the allotted time had expired. No specific attention was made to crafting a precise query, nor one that obtained high recall.

Due to the strict Boolean nature of PubMed, the number of results varied widely: 264 mean, 58 median, 367 standard deviation (number of results returned was capped at 1000). In Figure 6, we show the number of relevant documents retrieved in the top twenty (darker bars). The lighter bars represent the total number of retrieved documents if less than 20, which was the case for 16 topics. Due to these variations, P20 does not completely capture retrieval performance; for example, in the case where only one document was retrieved (and it was relevant), P20 would only be 0.05, yet the result cannot be any more precise.

Acknowledging that searchers vary widely in terms of information-seeking behavior, we are careful not to draw any generalizations from this particular set of queries. Instead, these results are intended

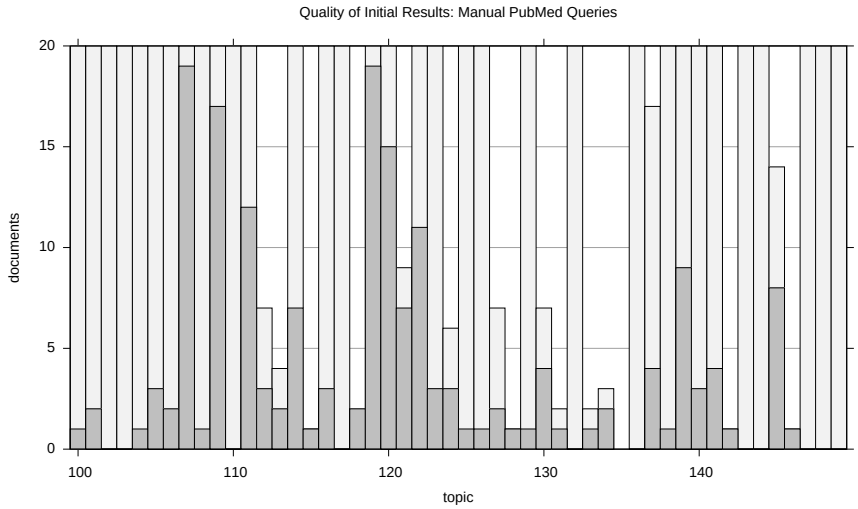

Figure 6: Schematic representation of the top 20 document retrieved by manual PubMed queries; relevant documents shown as darker bars.

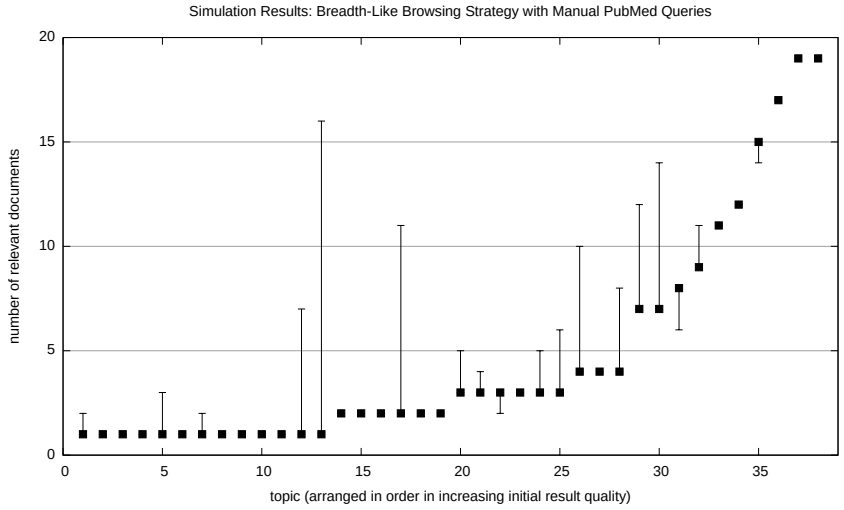

Figure 7: Results of applying breadth-like browsing strategy to manual queries. Solid markers represent initial results; bars represent simulation results.

to provide one example of how PubMed users might realistically behave. Our searcher might be characterized as an expert searcher in general (being an IR researcher), but one who is not specifically trained in the PubMed query language. We might imagine other classes of users who are domain experts and highly skilled in the interface (e.g., medical librarians), in which case higher performance would be expected. In these situations, there might be less of a need for browsing. Other users might be domain experts, but not specifically trained in search techniques (e.g., physicians), in which case performance might be comparable or lower. In these situations, browsing would certainly provide a useful tool.

We applied the breadth-like browsing strategy to the results of the manual PubMed queries. Simulated utility is shown in Figure 7. Instead of measuring P20, we counted the absolute number of relevant documents in the first 20 hits (which better accounts for cases where the initial result set contained fewer than 20 documents). Topics are arranged in ascending order of number of relevant documents in the initial results (shown as squares). The vertical bars denote the result of the simulations. For one topic, starting from a single relevant document, browsing related article suggestion would yield 15 additional relevant documents. The graph does not show the 11 topics for which the initial results contained zero relevant documents.

The simulation output appears consistent with the results in Section 5. On the whole, topics are difficult, and the initial results contain few relevant documents most of the time. Also, we see that topics with fewer relevant documents in the initial results get a bigger boost from browsing. Overall, this confirms the generality of our findings about the content-similarity browsing tool, independent of the actual search engine.

8 Conclusion

The contributions of this work are twofold: First, our experiments provided a deeper understanding of how PubMed users find information. We demonstrate that a browsing tool based on content similarity is able to compensate for poor retrieval quality, illustrating the complex relationship between retrieval performance and utility. Second, and more broadly, this work represents a case study in the application of user simulations for automatic utility evaluation. As a complement to both interactive user studies and Cranfield-styled experiments, simulation-based evaluations provide a powerful and flexible tool for formative studies.

Acknowledgements

For this work, the first author was supported in part by the National Library of Medicine. For this work, the second author was supported in part by the Center for Intelligent Information Retrieval and in part by the Defense Advanced Research Projects Agency (DARPA) under contract number HR0011-06-C-0023. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor. The first author wishes to thank Esther and Kiri for their kind support.

References

[1] I. Aalbersberg. Incremental relevance feedback. In *SIGIR 1992*.

[2] E. Agichtein, E. Brill, S. Dumais, and R. Ragno. Learning user interaction models for predicting Web search result preferences. In *SIGIR 2006*.

[3] J. Allan, B. Carterette, and J. Lewis. When will information retrieval be "good enough"? User effectiveness as a function of retrieval accuracy. In *SIGIR 2005*.

[4] A. Aula, P. Majaranta, and K.-J. Räihä. Eye-tracking reveals the personal styles for search result evaluation. In *INTERACT 2005*.

[5] E. Chi, P. Pirolli, K. Chen, and J. Pitkow. Using information scent to model user information needs and actions and the Web. In *CHI 2001*.

[6] C. Cleverdon, J. Mills, and E. Keen. Factors determining the performance of indexing systems. Two volumes, ASLIB Cranfield Research Project, Cranfield, England, 1968.

[7] W. Hersh, A. Cohen, J. Yang, R. Bhupatiraju, P. Roberts, and M. Hearst. TREC 2005 Genomics Track overview. In *TREC 2005*.

[8] W. Hersh and P. Over. Interactivity at the Text Retrieval Conference (TREC). *Information Processing and Management*, 37(3):365–367, 2001.

[9] W. Hersh, A. Turpin, S. Price, B. Chan, D. Kramer, L. Sacherek, and D. Olson. Do batch and user evaluations give the same results? In *SIGIR 2000*.

[10] M. Huggett and J. Lanir. Static reformulation: A user study of static hypertext for query-based reformulation. In *JCDL 2007*.

[11] M. Ivory and M. Hearst. The state of the art in automating usability evaluation of user interfaces. *ACM Computing Surveys*, 33(4):470–516, 2001.

[12] B. Jansen and A. Spink. How are we searching the World Wide Web? a comparison of nine search engine transaction logs. *Information Processing and Management*, 42(1):248–263, 2006.

[13] T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search. *ACM Transactions on Information Systems*, 25(2):1–27, 2007.

[14] K. Klöckner, N. Wirschum, and A. Jameson. Depth- and breadth-first processing of search result lists. In *CHI 2004, Extended Abstracts*.

[15] J. Koenemann and N. J. Belkin. A case for interaction: A study of interactive information retrieval behavior and effectiveness. In *CHI 1996*.

[16] Lemur. Lemur Toolkit for Language Modeling and IR, 2003.

[17] A. Leuski and J. Allan. Strategy-based interactive cluster visualization for information retrieval. *International Journal on Digital Libraries*, 3(2):170–184, 2000.

[18] J. Lin, M. DiCuccio, V. Grigoryan, and W. J. Wilbur. Exploring the effectiveness of related article search in PubMed. Technical report, University of Maryland, July 2007.

[19] J. Lin and W. J. Wilbur. PubMed related articles: A probabilistic topic-based model for content similarity. *BMC Bioinformatics*, 8:423, 2007.

[20] M. Magennis and C. J. van Rijsbergen. The potential and actual effectiveness of interactive query expansion. In *SIGIR 1997*.

[21] J. Mostafa, S. Mukhopadhyay, and M. Palakal. Simulation studies of different dimensions of users' interests and their impact on user modeling and information filtering. *Information Retrieval*, 6(2):199–223, 2003.

[22] M. Smucker and J. Allan. Find-Similar: Similarity browsing as a search tool. In *SIGIR 2006*.

[23] A. Turpin and W. R. Hersh. Why batch and user evaluations do no give the same results. In *SIGIR 2001*.

[24] A. Turpin and F. Scholer. User performance versus precision measures for simple search tasks. In *SIGIR 2006*.

[25] H. Turtle. Natural language vs. boolean query evaluation: A comparison of retrieval performance. In *SIGIR 1994*.

[26] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, London, 1979.

[27] R. White, I. Ruthven, J. Jose, and C. J. van Rijsbergen. Evaluating implicit feedback models using searcher simulations. *ACM Transactions on Information Systems*, 23(3):325–361, 2005.

[28] W. Wilbur and L. Coffee. The effectiveness of document neighboring in search enhancement. *Information Processing and Management*, 30(2):253–266, 1994.